DOCUMENT RESUME

ED 384 660 TM 023 954

AUTHOR Bejar, Isaac I.
TITLE A Generative Approach to Psychological and Educational Measurement.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-91-20
PUB DATE Mar 91
NOTE 58p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Educational Assessment; Item Response Theory; *Measurement Techniques; Models; Prediction; *Psychological Studies; *Psychometrics; Responses; Test Construction; *Test Items; Test Validity
IDENTIFIERS Generative Processes; *Response Generative Modeling

ABSTRACT

Response generative modeling (RGM) is an approach to psychological measurement that involves a "grammar" capable of assigning a psychometric description to every item in a universe of items and is capable of generating all the items in that universe. The article discusses the rationale behind RMG and its roots, explores how it relates to validity, and assesses its feasibility in a wide variety of domains. A brief review of possible theoretical approaches to a psychologically sound approach to test construction and modeling concludes the discussion. RGM links item construction and response modeling in a single package, so that linkage (the predictions about response behavior) is challenged every time a test is administered. The administration of a test then becomes a psychological experiment, a fact that may, in turn, lead to the improvement of both theories and tests. One table and seven figures illustrate the discussion. (Contains 133 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-91-20

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 384 660

A GENERATIVE APPROACH TO PSYCHOLOGICAL AND EDUCATIONAL MEASUREMENT

Isaac I. Bejar

BEST COPY AVAILABLE

Tm 023 954

Educational Testing Service
Princeton, New Jersey
March 1991

A Generative Approach to Psychological and Educational Measurement

Isaac I. Bejar

Educational Testing Service

[For publication in N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), Test Theory for a New Generation of Tests, Hillsdale, NJ: Lawrence Erlbaum Associates.]

Copyright © 1991. Educational Testing Service. All rights reserved.

Abstract

Response generative modeling (RGM) is an approach to psychological measurement which involves a "grammar" capable of assigning a psychometric description to every item in a universe of items *and* is also capable of generating all the items in that universe. The purpose of this chapter is to: 1) elaborate on the rationale behind RGM; 2) review its roots and how it relates to current thinking on validity; and 3) assess its feasibility in a wide variety of domains. The chapter concludes with a brief review of possible theoretical approaches to a psychologically sound approach to test construction and modelling.

A Generative Approach to Psychological and Educational Measurement

Introduction

Response generative modeling (RGM) is an approach to psychological measurement which involves a "grammar" capable of assigning a psychometric description to every item in a universe of items *and* is also capable of generating all the items in that universe (Bejar & Yocom, in press). Such an approach to measurement, if feasible, could have at least three important implications. First, the interpretation of scores from a generative instrument would be greatly facilitated because the process for generating the item is explicitly stated. Second, the possibility of generative modeling implies that we have a complete understanding of the underlying response process. Such knowledge might allow us, in turn, to abandon the multiple-choice format in favor of open-ended formats, a long-standing desire of psychometricians (e.g., Frederiksen, 1990) but without the expense associated with scoring open-ended responses. In other words, the same knowledge base that is used to create items can be brought to bear on the scoring of open-ended responses. Third, the ability to assign a *psychometric* description to an item is the key ingredient in what might be called *intelligent test development aids*. Job aids, in general, are rapidly becoming the key to increased productivity in many fields(e.g., Kline & Lester 1988; New York Times, 1989; Harmon, 1986). In a testing context, test development job aids might become essential if bills to outlaw pretesting succeed in becoming law, (because it is through pretesting that test developers estimate the difficulty of an item before the test is adminesterd in a final form) especially in light of growing statistical theory designed to allow equating tests "with little or no data." (Mislevy and Sheehan, 1990) Some speculations on the future of job aids for test development can be found in Bejar (1989); a discussion of open-ended assessment from a generative perspective, with special emphasis on certification testing can be found in Bejar (in preparation), see also Baker (1988)and the Summer 1989 issue of the *Journal of Educational Measurement*.

The purpose of this paper is to: 1) elaborate on the rationale behind RGM; 2) review its roots and how it relates to current thinking on validity; and 3) assess its feasibility in a wide variety of domains.

Historical Background

Although Item Response Theory (IRT) today enjoys unanimous endorsement of test developers and psychometricians, just some years ago other psychometric frameworks were serious contenders. One contender was Tryon's item sampling model (Tryon, 1957). He distinguished between three theories: the true-and-error-factor theory, which is a primitive IRT model; the theory of equivalent item samples, also known as a classic test theory (Gulliksen, 1950); and a theory based on random sampling from a universe of items, which Tryon endorsed. The tensions that lead to the item sampling model can be surmised from Osburn's (1968) influential paper:

> Few measurement specialists would quarrel with the premise that the fundamental objective of achievement testing is generalization. Yet the fact is that current procedures for the construction of achievement tests do not provide an unambiguous basis for generalization to a well defined universe of content. At worst, achievement tests consist of arbitrary collections of items thrown together in a haphazard manner. At best, such tests consist of items judged by subject matter experts to be relevant to and representative of some incompletely defined universe of content. In neither case can it be said that there is an unambiguous basis for generalization. *This is because the method of generating items and the criteria for the inclusion of items in the test cannot be stated in operational terms*. (p. 95; italics added)

Whereas local independence is the most critical assumption in IRT, the existence of a universe of items, or the possibility of generating one, was the core of the random sampling approach. And just as lack of local independence could prevent correct modelling of some abilities (e.g., Bock, Gibbon & Muraki, 1988, p. 277), an inability to formulate a universe of items could prevent the correct implementation of the random-sampling model. Loevinger (1965), for example, objected to the item sampling model because the

> term population [universe] implies that in principle one can catalog, or display, or index all possible members even though the population [universe] is infinite and the catalogue cannot be completed....No system is conceivable by which an index of all possible tests [items] could be drawn up. There is no *generating* principle (p. 147; italics added).

If Loevinger is correct then RGM would be doomed because RGM shares with the random sampling model the assumption that *there is* a generation principle. However, RGM does not require that the generated items constitute a random sample. Moreover, RGM goes much farther than the random sampling model by proposing that there is not only a generating but also that items be generated with psychometric parameters already estimated, as it were.

Strictly speaking, the random sampling model is a mathematical one, and by itself does not attempt to generate items. That component was to have been provided by an earlier attempt at generative item writing. The attempt that received most attention was that of Bormouth (1970), which was perceived at the time (e.g., Cronbach, 1970) as a potential breakthrough in item writing. However, the genesis of the approach appears to be in instructional psychology (e.g., Hively, 1974; Uttal, Rogers, Hieronymous & Pasich, 1970). An extensive summary of those efforts can be found in Roid and Haladyna (1982), a shorter one in Bejar (1983). The reason those efforts have not matured into a viable psychometric framework appears to be due to two factors: following too closely one source of inspiration, namely Chomskyan linguistics; and clinging to a behavioristic, as opposed to cognitive, orientation--in retrospect, quite paradoxical sources of inspiration.

Chomsky (1965) introduced the distinction between competence and performance to demarcate the purely linguistic phenomena from the psychological reality of language use. Competence refers to the universe of sentences that a user of the language *ought* to be able to comprehend or utter. In practice, of course, language users fail to comprehend certain sentences and make all kinds of grammatical mistakes when speaking or writing. Chomsky chose to focus on the phenomena of more linguistic relevance or "what the language user ought to know," rather than modelling actual language use, or performance. Both Bormouth and Hively also focused exclusively on

the competence and not the performance. That is, they aimed to generate the universe of items that students ought to be able to respond to correctly. This meant the generation of items without a concomitant psychometric description that might reflect the underlying response process required to respond to an item thus generated. The problem, as Merwin (1977) pointed out, was that what ought to have been the case often was not. For example, items generated to represent an educational objective were found to differ in their difficulty or the proportion of students who answered it correctly. There was no possible explanation for this variability in the absence of a performance component.

Interestingly, there were exemplars for the integration of competence and performance early on. Miller (1962), for example, proposed that the syntactic complexity of a given sentence would affect its comprehensibility, and called the theory the Derivational Theory of Complexity. The implicit performance model in the theory is that sentences require more, or less, mental computations depending on their syntactic attributes and therefore are harder, or easier to comprehend. That this approach was not recognized as a model for generative psychometrics may be in part due to the strong behavioristic trends in psychology and education at the time. It was, according to some historians (Gardner, 1985), Skinner's lack of rebuttal to Chomsky's (1965) critique of Skinner's (1957) *Verbal Behavior* that was the beginning of the end for behaviorism.[1]

In short, RGM shares some of the concerns with earlier attempts at generative modelling but in some respects could not be more different. Specifically, the item sampling model, and related item generation algorithms, constitute a psychometric model for classic behaviorists, for whom talk of underlying processes is not admissible. RGM, in contrast, has a cognitive orientation. This means that the postulation of underlying processes and knowledge structures required to respond to an item are not only admissible but at the heart of the approach: it is by incorporating information about the

[1] Of course, in psychology we can only speak of rounds. Behaviorism may be on its way back disguised as connectionism. Although behaviorism-as-connectionism opens the black box it might as well be kept closed: inspecting a neural net after it has been trained to emulate some human behavior is not likely to be informative, information is distributed throughout a network of nodes. Even when such a model accounts for verbal behavior (e.g., Rummelhart et al, 1986, but see Prince & Pinker, 1988) all we have learned, it seems, is that through pairing stimuli and responses learning can take place. The computational attractiveness of these models is undeniable, but it remains to be seen whether they will replace the computer as the metaphor to modelling human cognition. More likely connectionist ideas will be incorporated into cognitive models to improve the granularity of the account (Just, personal communication).

demands a given items imposes on the cognitive apparatus that it becomes possible to "pre-estimate" the parameters of some response model. Moreover, unlike the item sampling model, which rejects the postulation of latent ability, and therefore is philosophically at the other extreme of the IRT family of response models, RGM is compatible with IRT.

The scope of RGM is not limited to "achievement" items as, many of the earlier attempts to generative item writing were. As we will see below, RGM is, in principle, applicable to any domain, including achievement and instructional domains. In fact, a forerunner of the RGM can be found in an instructional context. Uttal et al. (1970) used the term generative instruction to describe an alternative to the machine learning efforts of the 60s, which were based on Skinnerian principles. The purpose of generative instruction is not to strengthen the linkage between a stimulus and a response but rather to diagnose the source of difficulties in learning. This idea was subsequently elaborated by Brown and Burton (1978) in the context of arithmetic instruction. In short, a generative approach cuts across domains and, as we will see, is a natural framework for the assessment of complex skills, such as troubleshooting, clinical diagnoses, and pedagogical skills.

RGM as an Approach to Validation

In addition to integrating the modeling of content and response, RGM exemplifies an approach to construct validation. Validation has traditionally focused on an accounting of *response consistency* or covariation among items. Indeed, construct validation has been described as implying "a joint convergent and discriminant strategy entailing both substantive coverage and response consistency in concert" (Messick, 1981, p. 575). There has been far less emphasis on an accounting of *response difficulty* (but see e.g., Campbell, 1961; Carroll, 1980; Davies, & Davies, 1965; Egan, 1979; Elithorn, Jones, Kerr, & Lee, 1964; Tate, 1948; Zimmerman, 1954). These two focuses, response consistency and response difficulty, are not antithetical by any means. Embretson (1983) has proposed an approach to validity in which both considerations are integrated. From this validational perspective knowing the latent structure of a test--for example, its factorial structure or its fit to a particular item response model--is clearly essential to an interpretation of test scores but is not the entire story. An accounting of response difficulty would clearly enhance the validational status of a test because to

obtain that accounting a model incorporating the mental structures and processes needed to solve the item would be required. If that model has been derived from a theory that has empirical support then, clearly, the validational status of the test scores derived from such a test have a head start, compared to a test developed following the actuarial model where the characteristics of the items are not known until it is administered to a sample of examinees.

Not only are accountings of response difficulty and consistency not antithetical, they entail parallel considerations. For example, within the response-consistency tradition, the extent to which covariation is accounted for by relevant and irrelevant (e.g., method) variables is often the basic data from which validity is assessed (e.g., Campbell & Fiske, 1959). A similar consideration is equally applicable in an accounting of response difficulty. For example, if it were shown to be the case that the difficulty of analogy items from, say, the SAT or the GRE were purely a function of word difficulty, then we could reasonably conclude that the validity of scores derived from such items would be suspect[2].

Psychological theorizing has changed substantially since the original article on construct validity (Cronbach & Meehl, 1955). The current strength of the cognitive perspective has led psychology from functionalistic theories to structuralist theories. More specifically, psychology now emphasizes explaining performance on the basis of the systems and subsystems of underlying processes and structures rather than identifying antecedent-consequent relationships. Cronbach and Meehl's emphasis on building theory through the nomological network, which contained primarily antecedent (test score) to consequent (other measures) relationships, can be viewed as a functionalistic approach.

Embretson (1983) has proposed a major reformulation of the validation process consisting of two stages: construct representation and nomothetic span (Embretson, 1983). This reformulation can be viewed as the culmination of debates on the role of structure and function in individual differences psychology (e.g., Messick, 1972; Carroll, 1972.)[3] In Embretson's reformulation, a construct is a

[2]Actually, with our increased understanding of the process of vocabulary acquisition (e.g., Sternberg, 1987; Curtis 1987) good performance on a vocabulary test can not really be discarded as an indication that the person is merely studious. Research suggests that vocabulary scores are good predictors of academic criteria because the process of vocabulary acquisition is a form of reasoning, which presumably accounts for the correlation of vocabulary tests with other tests.

[3] Structure and function are ambiguous terms. Messick (1971), for example associates *structure* with the results of factor

theoretical variable that is a source of individual differences *Construct-representation* research seeks to identify the theoretical mechanisms that underlie task performance by cognitive task analysis methods. That is, the component processes, strategies, and knowledge structures that underlie performance identify the construct(s) that is (are) involved in the task. *Nomothetic-span* research, in contrast, concerns the utility of the test for measuring individual differences. It refers to the span of relationships between the test score and other measures. Nomothetic span is supported by the frequency, magnitude, and pattern of relationships of the test score with other measures.

In Cronbach and Meehl's conceptualization, the correlations of individual differences on the test with other measures both define the construct and determine the quality of the test as a measure of individual differences. In Embretson's integrated conceptualization of construct, validity has qualitatively different types of data to support construct representation and nomothetic span. The former is supported by data on how *within-task variation* in the items' attributes influence performance, while the latter is supported by *between-task covariation*, for example, correlation among tests.

Summary. In short, RGM capitalizes on the convergence of several trends and can be seen as an approach to implement a structural perspective of validation by integrating item development, response model fitting, and validation. RGM integrates all three processes into a unified framework where item creation is guided by knowledge of psychology of the domain, and concomitantly psychometric descriptions (e.g., parameters on an IRT model) are attached to the item as it is generated. Then, every time a test is administered the psychology of the domain is tested, by contrasting the theoretical psychometric description with the performance of examinees, thus perennially assessing the validity of the scores. This approach to validation has much in common with other efforts to develop and validate psychologically-inspired tests or batteries (e.g., Frederiksen's (1986);Guttman (1969, 1980); Kyllonen (1990))

---

analysis, and talks about the *functional* links among traits and performance outcomes. Guttman (1971), however, associates *structure* with the system of a priori relations among variables (see Lohman and Ippel, this volume). The term construct representation in Embretson's formulation has both structural and functional overtones, whereas nomological span, which coincides with Cronbach and Meehl (1955) nomological network idea, is primarily functional.

BEST COPY AVAILABLE

Evidence for the Feasibility of RGM

The two major ingredients for a generative approach are (1) a mechanism for generating items and (2) sufficient knowledge about the response process to estimate the psychometric parameters of the generated items. The feasibility of the approach, therefore, can be judged by whether items can, in fact, be generated and whether the predicted parameters are, in fact, observed. In the following sections I will present evidence, from my own research and that of others, suggesting that RGM is indeed feasible. At times, however, the discussion will turn speculative because in some domains where the approach would seem feasible no attempts to implement generative modelling have been made.

Spatial Ability

Not surprisingly, good examples of the feasibility of RGM can be found in the domain of spatial ability. For one thing, the generation of spatial items seems simpler, for another spatial ability has been under intense scrutiny of cognitive psychologists. In this section I present evidence for mental rotation items and hidden figure items (see also Irvine, Dunn & Anderson, 1989).

Mental rotation. It is seldom the case that sufficient knowledge has accumulated about an ability to make RGM immediately feasible. One exception is mental rotation. Although psychometricians have long used two-dimensional figural rotations in tests, it was experimental psychologists (Shepard & Metzler, 1971) who thoroughly analyzed the mental process. There now exists a large body of literature (cf. Corballis, 1982) establishing that an angular disparity between the two figures largely determines the time to respond.

A generative approach to the measurement of this ability means controlling the difficulty of an item through the angular disparity between two stimuli. Imagine, for example, a test consisting of, say, 20 distinct pairs of figures which can be presented at rotations ranging from 20 to 180 degrees. In an adaptive test every examinee would be presented with the 20 items, but examinees of different levels of ability would be presented with items at a different angle. Clearly, such an adaptive procedure requires a computer. All examinees would perhaps be given the first pair at 100 degrees. A higher ability examinee would then be presented subsequent items at larger rotations. Although it might be feasible

to tailor the test to the examinee and score on the basis of rotation angle alone, in practice there are at least two problems with that idea. First, the difficulty of any given item is a function of not only rotation but also the complexity of the figure. Second, mental rotation is the type of skill where speed of response is an appropriate consideration. Therefore, in order to use all the information we need to calibrate each item separately and record how long it takes the examinee to respond.

To judge the feasibility of RGM for this task requires that we calibrate several pairs of figures on some item response model and that we estimate the difficulty of the pair at several degrees of rotation. The expectation for mental rotation data is that the relationship of difficulty on angular rotation is linear for several elapsed times (Bejar, in press). The expectation was tested by fitting the simplest possible psychometric model of an 80-item test based on figures such as those in Figure 1. The examinee's task is to determine if the figure on the right is a rotation of the one on the left. There were eight basic items presented at five angles (20, 60, 100, 140, and 180 degrees) in their true-and-false version (in the false version the second figure is the mirror image of the first figure), in order to establish the relationship between angular disparity and difficulty.

Figure 1: Sample mental rotation item

Figure 2 shows the result of a calibration for a typical item based on the responses of nearly 200 high school students. As can be seen, there are some departures from the predictions although, in general, the fit for this item is good. The major deviation from linearity occurred at 100 degrees. Also, beyond 5 seconds a tendency towards a quadratic relationship between difficulty and angular disparity emerges, a situation which suggests that beyond a certain elapsed time different response strategies may come into play. In principle, such departures from linearity might be avoided by adapting the test to the examinee, which was not done with these data. In other words, so long as the item is not too difficult for an examinee responses may in fact be just a function of angular disparity.

The results for the false items are quite different in that angular disparity does not seem to control response time, as it does for the true items. That is, the false items seem to tap the decision

aspect of performance, while the true items are tapping the mental rotation aspect. Needless to say, this introduces a complication. Thus, it may not be practical to use a true-false format in a real application. A multiple-choice version may eliminate the problem but introduces the complexity that the attributes of the alternatives would have to be considered in the modeling process.

Figure 2:Relationship of estimated difficulty on angular disparity at several elapsed times

---

Hidden figure items. Unlike the mental-rotation items, for which the determinants of performance are understood, very little is known about the determinants of performance on hidden-figure items. A theory that addresses performance on tasks of this type has been proposed by Duncan and Humphreys (1989) and although it was not used as the inspiration for representing hidden figure items, it is consistent with the representation that was chosen. That representation needs to capture not only the complexity of the item but also lend itself to generating items that have the same underlying representation but a different visual realization, that is items that should have the same difficulty but appear visually different. For convenience, we call the items generated in this fashion *clones*, although they could also be called *isomorphs*, as is done by some cognitive researchers interested in the cognitive equivalence of problems (e.g., Kotovsky & Simon, 1988). Figure 3 shows a typical hidden figure item and a corresponding clone. The task for the examinee is to determine if the smaller figure is embedded in the larger one.

Figure 3: Typical true hidden-figure item and two corresponding clones

---

The representation chosen to represent items and obtain clones was a matrix consisting of counts indicating how close the target figure appears at each possible position in the larger pattern and was based on the Hough transform (Mayhew & Frisby, 1984), an artificial intelligence technique used in object recognition (see Bejar & Yocom, in press). We tested the validity of this representation by implementing a computer program capable of generating clones and then, comparing their psychometric characteristics on the basis of responses from high school students. In other words, we tested the psychometric equivalence of pairs of isomorphs or clones. This "weakened" version of full

generative modelling, where instead of generating items of known difficulty we just generate items that have the same difficulty as the generating item, was necessary because the lack of theoretical development for performance on this item type. The results demonstrated that the clones behaved as such in terms of their difficulty as well as distribution of response times. Figure 4 shows the relationship between the logit for proportion correct for pairs of clones as well as the corresponding mean response time. Figure 5 shows the cumulative response times for two clones. It can be seen they are very similar, and this was true for the other items as well.

Figure 4:Regression of logit of proportion correct for pairs of clones (a) and the corresponding mean response time
Figure 5:Cumulative response time for a pair of clones

Reasoning Tests

Reasoning tests, both deductive and inductive, lend themselves to generative modeling. In this section we discuss the impressive evidence for inductive reasoning provided by Butterfield, Nielsen, Tangen and Richardson (1985) using letter series, preliminary evidence on analogical reasoning, and speculate on the feasibility of generative modeling of deductive and quantitative reasoning items.

Inductive reasoning. Butterfield et al. describe a comprehensive approach to describing and generating letter series, as well as a theory of item difficulty, the two ingredients of generative modeling. The items consist of series of letters produced according to a set of rules and the examinees task is to predict the next element in the series. Arbitrary series can be generated by applying operators to generate the next letter in the series. The operators considered by Butterfield et al. are Next (N), Back (B) and Identical (I). The generic form of an item can then be succinctly described as the rules of construction in terms of these operators.

The following item,

DDQQEEPPFFOO

is described by the form $N_1I_1B_2I_2$ and two starting values, in this case C and S. The subscripts refer to

a position in the starting string. From a starting value of C we first apply $N_1(C) = D$, yielding a D, and then apply $I_1(D) = D$, yielding another D. We now move to the second element of the string which starts from S. Applying the B operator yields $B_2(S) = Q$ and applying I yields $I_2(Q) = Q$. In short, Butterfield at al. are able to characterize abstractly series as well as generating series that have a given abstract characterization. Although oriented to open-ended series their methodology can be used with multiple-choice versions as well. The following multiple-choice item from the Factor-Referenced Cognitive Tests Kit (Ekstrom, French & Harman, 1976) asks the examinee to choose the series that does not belong.

NOPQ DEFK ABCD HIJK UVWX

The first, third, fourth and fifth can be represented by the rule N_1 with starting points N A, H, and V respectively. Thus, to create multiple-choice versions one would use the theory to generate options that have the same generation principle.

In addition to characterizing items abstractly, RGM requires a mapping from that characterization to the parameters of psychometric model, such as difficulty. Butterfield at al., building upon earlier research by Simon and Kotovsky (1963), proposed and demonstrated a theory of item difficulty that suggests that the difficulty of a series is indexed by the knowledge required to discover the most-difficult-to-represent string in the series. They also propose several indices of that representational difficulty. Several experiments demonstrated the validity of the scheme. Moreover, when applied to predict the difficulty of items in the Primary Mental Abilities Test they accounted for 90% of the variance in item difficulty. This is impressive because those items did not enter into the formulation of the theory.

Deductive reasoning. There is not really a comprehensive demonstration of RGM deductive reasoning. There are, however, several lines of research concerned with among other things an accounting of difficulty of several types of deductive reasoning tasks. This accumulation of results and variety of theoretical accounts (see Galotti, 1989) would make it an excellent domain for attempting a

generative approach. Moreover, because of the conflicting accounts of deductive reasoning such an investigation may have psychometric value as well as helping to shed some light on the field.

The work of Johnson-Laird, Byrne and Tabossi (1989) illustrates the potential feasibility based on a mental models approach. A mental model consists of "tokens arranged in a particular structure to represent a state of affairs" (Johnson-Laird, 1983, p. 398). Specifically, Johnson-Laird at al., propose and show that the difficulty of problems with multiply quantified premises, e.g.,

> None of the Princeton letters are in the same place as any of the Cambridge letters.
>
> All the Cambridge letters are in the same place as all the Dublin letters.
>
> Therefore, none of the Princeton letters are in the same place as any of the Dublin letters.

They show that the difficulty of the problems is a function of the number of the mental models that the solver needs to postulate to solve the problem: Problems that required a single model were found to be easier than problems that required two mental models. A theory of difficulty that accounts for only two levels of difficulty has a long way to go for psychometric purposes. On the other hand, the generation of deductive reasoning items would not present serious difficulties because of their rigid format. In short, generative modelling of deductive problem solving appears feasible, but further work is needed to fully account for variations in difficulty. A complete accounting will require incorporation of biases that test takers follow when asked to think deductively. An approach that is generative in spirit but incorporate logical biases in item construction has been described by Colberg and Nester (1987).

Analogical reasoning. Analogical problem solving has a long psychometric tradition but surprisingly little is known about the formal characteristics of such items. A recent study (Bejar, Chaffin and Embretson, in press) has begun to remedy the situation by studying intensely a large number of analogy items from the Graduate Record Examination (GRE) General Test. The study showed that despite the fact that the analogies are in a verbal modality, vocabulary knowledge, as such, is not even remotely the main determinant of performance or item difficulty. (Of course, vocabulary

knowledge is required to answer the items but the more difficult items are not so because they involve infrequent words.)

The generation of analogies has been demonstrated by Chaffin and Hermann (1987). The possibility of generative modeling of analogical reasoning, that is, generating items with known psychometric characteristics was considered by Bejar at al. (in press). They concluded that given the current state of the art in computational linguistics, working at the word pair level was more feasible. By using word-pairs as the building block multiple-choice generative modeling could be implemented in this fashion: Prepare a database of word pairs and store along with the word pair information such as the semantic relational features of the word pair, the frequency of the words making the pair and possibly other information as well. The generation of an item starts by deciding which major semantic class to use. Bejar at al., found 10 major classes in the GRE item pool. Each major class has distinctive features that, in turn, makes it possible to classify word pairs into subclasses. Thus, to create an item we chose the stem and the key to be from the same subclass and chose options that are from the same class but different subclasses. Thus, the template for creating analogy items is:

Stem: Word-pair$_{ij}$

Key:Word-pair$_{ij}$, where i = j

Nonkey:Word-pair$_{ij}$, where i < > j,

where i refers to a major semantic class, such as part-whole, class-inclusion, etc; j refers to a subclass within the major class. Essentially the template says that the stem and the key should be from the same class and subclass whereas the non-keys should be from the same major class but different subclasses. Clearly, this approach assumes that a semantic analysis is available for each word pair in our database, a process which at the moment must be done "by hand" (but see Miller, Fellbaum, Kegl & Miller, 1988; Byrd, Calzolari, Chodorow, Edwards, Klavans & Neff, 1987 for advances in computational linguistics that may eventually allow an automated implementation).

Constructing items according to a semantic analysis would qualify as generative were it not for

the fact that the semantic class is a potent determinant of difficulty. Bejar at al. studied different factors of difficulty and found that for the GRE pool the semantic class was the strongest determinant and not word frequency as Carroll (1980) had speculated, nor processing demands as we would have expected from recent research (Sternberg, 1977; Pellegrino & Glaser 1982).

Although the difficulties of generating multiple-choice analogies does not appear insurmountable, it may be easier to do so in an open-ended format. The first idea that comes to mind for an open-ended analogy item is to present the examinee with a word pair and then ask the examinee to produce one or more pairs that exemplify the same relation. This approach, however, is not likely to be adequate because the granularity of a typical multiple-choice item is very fine and therefore require responses that demand a high level of reasoning. That is, the exact nature of the relation represented by the stem is not certain until the options are examined. For example, a stem like grain:husk obviously calls for a part-whole relationship, but in the context of a GRE or SAT item the options would all be part-whole relationships, which requires the examinee to determine the exact kind of part whole relationship.

A format that preserves the inductive nature in an open-ended format is the analogical series, where the stem consists of two or more word pairs that specify the nature of intended analogy. We will discuss it briefly to illustrate the claim made earlier, namely that the knowledge that makes possible generative modelling may make it possible to abandon the multiple-choice format in favor of open-ended items.

Consider the following analogical series where the examinee is asked to provide one or more word pairs consistent with the series:

husk:grain, shell:turtle

The solution is not just any part-whole word pair but one where the part plays a protective function. A possible correct answer is armour:knight or peel:orange. This format is compatible with recent theorizing about the nature of analogical reasoning. Earlier theories focused almost exclusively on

processing models and paid no attention to the structure of knowledge. More recent theorizing (eg., Gentner, 1983) by contrast emphasizes the structural details of the process.

In short, a generative approach to either multiple-choice and open-ended analogical reasoning based on word pairs as the "building blocks" seems feasible because of advances in our understanding of performance on such tasks, such as the role of the semantic class on difficulty, and improvements in our understanding of the nature of the analogical process itself (e.g., Gentner, 1983), and advances in computational linguistics.

Quantitative and arithmetic reasoning. As one might have suspected, arithmetic and quantitative items lends themselves well to a generative approach. It is not difficult to think in the case of arithmetic, for example, of means of generating items (see Roid and Haladyna, 1982). For the same reason, the factors that might affect difficulty naturally suggest themselves. The most prominent line of research on difficulty factors is called "task variables". The culmination of this line of research can be found in the volume edited by Goldin and McClintock (1984).

The work on automated generation of quantitative items, however, has evolved independently of the work on task variables and for the most part has concentrated on arithmetic problems (e.g., Hively, Paterson & Page, 1968). However, it also ignored psychometric difficulty as an attribute of the generated items (see Merwin, 1977). As a result of this lack of convergence between research on determinants of difficulty and item generation we cannot point to an exemplar of generative modeling of arithmetic or quantitative reasoning. However, implicit in Brown and Burton (1978) work on diagnosis of arithmetic skills there is a problem generation mechanism that aims to generate items that would be consistent with the current diagnosis (see Burton, 1982) and illustrates that generative modelling need not be associated with a specific measurement framework, such as IRT. In a diagnostic context the questions to be administered next should be those that are most informative with respect to the different diagnoses under consideration. Obviously, this purpose of measurement calls for a different representation of the examinee. We will discuss some of these representations below under a discussion of the assessment of complex skills.

Quantitative skills involve more than arithmetic computations, of course. The solution of word

problems is perhaps a more important component of quantitative reasoning. Much of the early work on word problems focused on surface variables of the problem, or at least on a characterization of the problem without necessarily establishing that such characterization in any way was consistent with the problem as approached by the examinee. An important chapter by Riley, Greeno, and Heller (1983) may have changed that. They distinguished between the "specific" and "global factors" that affect problem difficulty. Global factors refer to surface characteristics of the problem. Specific factors refer to the deep characteristics of the problem which describes the relationships among the quantities involved in the problem. The taxonomy of specific factors they proposed consisted of four classifications: Change, Equalize, Combine and Compare. Each of these types has a schema associated with it that embodies the understanding required for solving problems of that type.

Another approach to classifying quantitative reasoning problems has been provided by S. K. Reed (e.g., Reed, Ackinclose & Voss, 1990), who categorizes problems into classes, such as Cost, Distance, Fulcrum, Work, etc., and then within each such class by the equation implied by the problem. For example, the following is a Cost problem:

> A group of people paid $238 to purchase tickets to a play. How many people were in the group if the tickets cost $14 each? (Reed at al., p. 85)

The equation that characterizes this problem is $14 = $238/n. Although the classification has been found useful for tutoring purposes, for generative modelling purposes further detail would be needed. In the above problem there are three quantities involved: the number of people, the cost of the ticket, and the total price. Therefore, variants of the above problem are possible as follows:

> Ten people paid $238 to purchase tickets to a play. How much did they pay for each ticket?
>
> Ten people went to see a play and each paid $14 per ticket. How much did they pay altogether?

In general, given n variables there will be n problem-variants, if we limit our attention to considering quantities as given or unknown. In reality, there are more variants because the quantities involved in the problem can enter into different types of relations. For example, in motion problems the entities may be traveling in the same or opposite directions. We refer the reader to the important work of Hall, Kibler, Wenger, and Truxaw (1989) and Mayer (1981), who seem to have provided, so far, the most comprehensive taxonomies of quantitative reasoning items.

With these taxonomies in hand, the generative modelling of quantitative reasoning might proceed by estimating the difficulty of items in cells of a multidimensional taxonomy. The generation of items from a given cell would necessarily be based on templates or well-defined scripts from which specific isomorphs could be generated. The validation of the generation of items from cells in this taxonomy could be assessed by the degree to which the psychometric parameters from a given cell are well-predicted and the within-cell residuals are constant across all cells. Unless the latter holds there are performance factors that are not captured by the taxonomy and the generative modelling is not complete. Stating generative modelling in this form makes it evident that methods derived from generalizability theory have relevance to RGM when we focus on the item as the unit of study, instead of the examinee. Specifically, methods for test constructed from tables of specifications (Jarjoura & Brennan, 1982; Kolen & Harris, 1987) seem relevant.

Verbal Ability

Verbal ability is measured by tasks such as sentence completion, reading comprehension and vocabulary tests. Vocabulary tests, despite their simplicity, are one of the best predictors of intelligence (Sternberg, 1987). The high correlation between intelligence and performance on a vocabulary test has been a bit of a mystery, but as a result of research on the nature of vocabulary acquisition it is now clear that the reason for the correlation was that performance on vocabulary tests is an indicator of the knowledge acquisition ability of the examinee (Jensen, 1980, p. 146).

Vocabulary. The generation of multiple-choice vocabulary tests by computer would appear to be trivial We might choose two synonyms to play the stem and key roles and then choose other words

for the distractors. Examination of vocabulary tests, however, reveals that the distractors are chosen in such a way that they are not unrelated to the stem. Therefore, difficulty is to some extent a function of the likelihood that the examinee has encountered the words included in the item but also how close the distractors are to the stem. As items get more difficult, the examinee must make finer distinctions. Therefore, in order to generate items of a wide range of difficulty the generation procedure would have to have access to a finely-tuned lexical database. Psychologically-motivated lexical databases are not readily available at the moment but may be in the future (Miller at al., 1988) and at the very least would be useful to assist the test developer in constructing items.

Interestingly, the measurement of verbal ability through sentence-based items appears more immediately feasible. Bejar (1988) discussed a system for the assessment of writing ability, which could easily be applied to sentence completion as well. The system relied on a grammar correction engine known as WordMAP published by Linguistic Technologies. The system envisioned by Bejar (1988) is shown in Figure 6.

Figure 6:System for generative assessment with sentence-based items.

It assumes a database of sentences from which items would be created. The system does not aim to generate natural text but rather to generate items based on sentences that have been previously selected for their suitability to assess specific writing errors. Because performance would be expected to depend on a variety of syntactic and semantic attributes of the sentence (e.g., Bejar, Stabler & Camp, 1987) that information would be stored along with the sentence.

The system would generate an item by choosing a sentence from the database and introducing an error, for example, a subject-verb agreement. The sentence with the error is then presented to the examinee who would rewrite it to remove the error. Scoring of the corrected sentence is possible through a "grammar engine". Bejar (1988) showed that WordMap could handle most of the constructions and errors in the Test of Standard Written English (TSWE). More recently, Breland and Lytle (1990) showed that WordMap could be used to score actual essays. That is, counts obtained from WordMap regarding errors and style were shown to predict ratings from readers very well. WordMap

has no idea about the meaning of the text it analyzes, but the results from Breland and Lytle suggest that it can be used in lieu of a second rater.

Ackerman and Smith (1988) has shown that measurement of writing ability should include both sentence mechanics and essays. The results presented in this section suggest that generative sentence-based assessment of sentence mechanics could be coupled with computer-scored essays and the score from a single rater into a more valid but less expensive measure of writing ability.

Reading comprehension. Reading comprehension, as measured by sentence completion items could be implemented generatively with a system similar to the one in Figure 6, except that instead of introducing a grammatical or stylistic flaw into the sentence a word would be omitted. Unfortunately, very little is known about the sentence completion item type despite the fact that is used by most admissions test. Examination of a number of these items suggests that not any sentence lends itself to be a stem for a sentence completion item and that a small set of rules would account for the choice of deletion (Fellbaum, 1987).

The assessment of reading comprehension through the reading of longer texts takes two forms. One is based on the cloze procedure, where words are deleted from the text according to a set of rules, and the examinee is supposed to replace the word, or choose from a set of possible replacements. The other possibility for measuring reading comprehension, found in most admissions tests, is to present a text and then ask questions about the text. Generative modelling for this item type would seem to especially challenging. First, it requires an understanding of the effect of text attributes on comprehension and secondly a procedure to generate questions about the text.

A characteristic of typical items of this type is that performance, as in most reading tasks (Just & Carpenter, 1987), requires background knowledge. That is, reading comprehension is a function of the attributes of the text but also of what the examinee brings to the reading task. In fact, Perfetti (1989) has distinguished between reading *comprehension* and reading *interpretation* to emphasize that what he calls interpretation requires both extracting the meaning from the text and applying world knowledge to it, whereas what he calls comprehension comprehension is just extracting the meaning from the text. Generative modelling of interpretation appears especially challenging because in effect

the question generation mechanism would have to have world knowledge equivalent to that of potential examinees. On the other hand, generative modelling of what Perfetti calls comprehension requires a mechanism for posing questions based on a given text and a theory of difficulty to anticipate the difficulty of those questions. Katz (1988) has developed a system called START which automatically analyzes English test and automatically transforms it into a propositional representation in such a way that questions based on the text can be generated. Examination of the questions generated by START for a GRE passage show, however, that they are of the factual type, and would not be appropriate for the measurement of reading ability of prospective graduate students. Nevertheless, the system might have applications for younger testees and in the assessment of English as a second language, if a theory of difficulty can be developed for it.

In short, generative modelling of reading comprehension appears especially difficult because the role background knowledge plays on performance and because questions that best tap that comprehension must call on background knowledge as well as the specifics of the text.

Complex Skills

In this section I discuss the assessment of skills that are not well characterized by a total score and call for a richer representation of the tasks and the examinee. First, I discuss achievement testing of the type that takes place in computer-based instruction where the computer would, ideally, guide the student through an optimal path. Next, I discuss the assessment of pedagogic skills. Finally, I discuss generative assessment of trouble shooting and diagnostic assessment skills.

Achievement testing. A generative approach to achievement testing remains to be developed. Part of the challenge no doubt is due to the elusive nature of the concept of achievement (cf. Green, 1974; Cole, 1990). A generative approach that is consonant with current thinking on the nature of learning (e.g., Glaser, 1988) is likely to be different from the approaches we have discussed for the assessment of generic abilities because ranking individuals would not be the focus of measurement. In achievement testing we are often interested in providing diagnostic information for a student, a teacher, or a computer to formulate an instructional plan. Therefore, the selection of questions would not be based on difficulty, but rather on the degree of information that the answer to a question would provide

in updating the several hypotheses under consideration to account for a student state of knowledge. An example of this approach is illustrated by the work on fractions of Brown and Burton (1978). The essence of the approach is to concoct the next item so that it would be maximally informative with respect to a hypothesis about the misconceptions harbored by a student. Although their notion of explaining performance in terms of bugs is not currently widely endorsed by cognitive psychologists, the general approach remains sound (e.g., Bejar, 1984) and has even been cast within an IRT framework (Tatsuoka & Tatsuoka, 1987; Yamamoto, this volume).

In general, achievement testing that is also diagnostic requires that we represent a student not as a point on a scale but rather as a complex data structure, such as a vector of misconceptions or a network, the nodes of which could stand for beliefs, hypothesis, concepts, etc, that describe the student's knowledge state. The purpose of measurement then is to estimate the activation, i.e., the degree to which concepts and beliefs, for example, are present, as well as the interconnectedness among the concepts. Traditional measurement models are not oriented to representing the examinee in that form and therefore a methodology is lacking for estimating achievement for such complex representations of the student. Although such representations are the essence of cognitive models, utilizing them for measurement, rather than description, is not common yet. A description is a declaration or set of assertions about the knowledge state of a student without inferential power. Measurement, by contrast entails generalizations, given a description. For example, given an ability estimate based on an IRT model we can make inferences about the probability of that someone with that ability will respond to other items measuring the same ability. Thus, for cognitive descriptions to qualify as measures we need to be able to estimate them and demonstrate their inferential power (cf Mislevy, this volume).

The advent of connectionist computational models opens up interesting possibilities because of the flexibility they provide to model a wide variety of phenomena as well as for their computational convenience. As an example, consider the modeling of physics knowledge in terms of beliefs about physical observations (Ranney & Thagard, 1988). In this case the description consists of a network of nodes for a given student. Some of these nodes stand for evidence, world knowledge, hypotheses, and

explanations that describe the student's knowledge state. Ranney and Thagard (1988) build the network by transcribing a think-aloud protocol into nodes and connections among nodes. What makes their system suitable as a measurement tool is that they superimpose a set of constraints on the network, based on Thagard's theory of explanatory coherence (Thagard, 1989). For example, among the principles or constraints proposed by the theory are the analogy principle, which states that analogous hypotheses explaining analogous evidence are coherent with each other. These constraints have the effect of controlling the propagation of activations throughout the network. After each piece of new information the network is allowed "to settle." The settled network is then the current estimate of the student's knowledge state. A further characteristic of the approach that makes it suitable for assessment purposes is that the representation of the student as a network is dynamic. That is, as new information becomes available it can be propagated throughout the network. Thus, the network represents the state of knowledge or beliefs on a moment by moment basis.

An obstacle to becoming a practical method of assessment is the reliance of think-aloud protocols as a means of computing the initial network. However, it would seem feasible to bootstrap the network from a structured questioning procedure. That is, instead of expecting the student to verbalize observations and hypothesis through a think-aloud protocol, a questioning procedure would extract information from the student. Once the network is bootstrapped, predictions can be made about the student beliefs and tested against questions posed to assess those beliefs. The answer to each such question is further data to be fed to the network. The goal of the entire procedure is to move the student toward some ideal network. Therefore, the questioning procedure would have access not only to the student's network but also to a network representing an ideal student. Marshall (1990) has devised a related procedure for mathematic word problems. She presents a series of problems to a student and, after the student has worked a set of problems, responds to a structured questioning procedure about the problems just solved. The result is a network, which, at the moment, is used for descriptive purposes but could easily be used as the basis for dynamic instruction and assessment.

Teaching skills. Because generative modelling is based on a model of the examinee it has the potential to be used for the assessment of teachers as well. For example, the information used to

model the examinee can also be used "in reverse" to generate case studies for a teacher to diagnose. This would correspond with the generation of medical and troubleshooting scenarios to be discussed below. Such an approach to the assessment of teachers would be very much in line with the preoccupation of integrating an "expanding body of knowledge on children's learning and problem solving to classroom instruction" (Carpenter, Fennema, Peterson, Chiang & Loef, 1989, p. 500).

As with the characterization of expertise in other fields (e.g., Chi, Feltovich & Glaser, 1981), a cognitive approach has become fashionable (cf Borko & Livingston, 1989). For example, Borko and Livingston suggest that a characteristic of more experienced teachers is the ability to reason pedagogically, which means the ability of the teacher to adapt content knowledge to the background of a specific group of students (Shulman, 1987). Such reasoning presupposes the ability of the teacher to characterize, in some detail, each student's knowledge state. In other words, more experienced teachers are able "to predict misconceptions students may have and areas of learning these misconceptions are likely to affect" (Borko & Livingston, 1989, p. 491).

In short, the picture that emerges is that teacher expertise requires not only subject matter knowledge, which can be measured in the usual manner, but also the ability to transform that knowledge in such a way that students varying in their knowledge can benefit most effectively. Measures of the latter remain to be developed. One possibility is an assessment task that requires the candidate to characterize the knowledge state of a group of students. As part of the exercise the teacher would prepare a set of problems and simulate its administration to a group of students. The simulation would then return to the teacher the answers provided by each student. The teacher's task would then be to characterize each student's knowledge state. From there the simulation could continue in a number of directions. For example, as a next step the teacher might be asked to prepare a teaching plan that is suited to the mix of students generated by the simulation.

Troubleshooting. Tasks which require diagnostic expertise, such as equipment troubleshooting and clinical diagnosis, are naturals for generative assessment, especially if approached from a model-based perspective. For example, in a troubleshooting situation a model-based approach would estimate the mental representation of the device under consideration, i.e., the structural and

functional description of the device as known to the examinee (e.g., Kieras, 1990). This sense of model-based is seen in AI research to distinguish between model-based (deep) and shallow (rule-based) expert systems.

Table 1:Trouble shooting table

In short, a generative approach to the assessment of troubleshooting skills would be to infer the examinee's conception of the device from responses to short questions which tap knowledge of different aspects of the device. The tasks would be generated from an algorithm that has access to a description of the device and generates troubleshooting tasks that collectively tap all the procedural and device knowledge. An alternative approach is to present open-ended tasks and record all the actions taken by the examinee and infer from those actions their mental model of the device, as well as procedural, declarative, and heuristic knowledge. Both approaches are compatible because knowledge of the domain is required to generate discrete items and interpret open-ended performance . However, assessment based on short questions items may be more efficient without sacrificing information.

For example, consider the generative assessment of troubleshooting of the circuit in Figure 7. The circuit is a full adder after Fulton and Pepe (1990). The circuit has three commands that can be sent to the circuit and five responses (or measurements) that can be obtained from it. Table 1 shows the relationship between the 8 possible input configurations and the correct outputs. There are however, 32, possible output vectors (the number of distinct vectors of length n is, in general, 2^n or 2^5 in this case), which leave 27 possible troubleshooting tasks. Obviously, if the examinee can correctly pinpoint the problem in each of these 27 tasks he or she must have an adequate mental model of the device. The more interesting question is to infer the partial device in the examinee's mind when there is less than perfect performance.

Insert Figure 7:Electronic device

In practice, the assessment of troubleshooting skills is most likely to take place in an instructional context. Lesgold, Ivill-Friel and Bonar (1989) discuss a system for teaching basic electricity principles, where the system needs to know not only electricity but also must contain instructional expertise to guide instruction and testing.

As with device troubleshooting, the representation of medical expertise with "shallow," i.e., if-then rules, has been found to be inadequate for many purposes. Causal or model-based representations have now been proposed which have important uses in expert systems and for clinical training. An important by-product of that trend for assessment for a generative perspective is the possibility of generating clinical scenarios or patients. (e.g., Parker & Miller, 1988; Miller, 1984; Pearl, 1987, Chapter 4). When a clinical scenario is represented as a probabilistic causal network it is possible to update the network as new information becomes available, from, say, clinical tests ordered by the examinee, or other simulated clinician-patient interactions. Actions and decisions can then be evaluated with respect to a perfect clinician represented by the network. Some ideas for generative assessment of medical expertise are discussed by Braun, Carlson and Bejar (1989). A system that lends itself to measurement from that perspective has been discussed by Warner and associates (1988).

Conclusions

There is a growing concern among some psychometricians (e.g., Goldstein & Wood, 1989) that the kind of theorizing that accompanies Item Response *Theory* has little to do with what the test it is applied to is supposed to measure. They even suggest that the research performed under the IRT rubric should be relabelled Item Response *Modelling* instead. This paper is, in a sense, a constructive reaction to the concern and evolves naturally from attempts within the IRT tradition (e.g., Fischer 1973) to incorporate substantive or collateral detail as part of the response modelling process. It also represents an example of what Snow and Lohman (1988) call the link between laboratory and field. RGM not only links laboratory and field but also challenges the item writer and psychometrician to test their knowledge base constantly, indeed every time a test is administered.

While the foregoing results point to the feasibility of an approach to measurement where response modelling and response theory are integrated under a generative framework, it also raises the question of whether there is a single psychological framework under which such an ambitious undertaking would fit. Even if RGM were successfully implemented in a wide range of domains chances are that somewhat incompatible procedures and assumptions would be used to model each domain. This is because RGM is not a psychological theory, or even a methodology, but rather a philosophy of test construction and response modelling that calls for their integration. It is more than likely that the application of RGM to specific item types will not yield a coherent picture that encompasses a multitude of domains. A complete picture requires an account of *inter-domain* covariation, that is the relationship of test performance across different domains, as well as *within-domain* variation in item parameters. The challenge, therefore, would be to model specific domains through a common set of assumptions in such a way that the within-domain psychometric characteristics can be anticipated as well as inter-domain covariation.

Stating the challenge in this form underscores the communality that exists between cognitive psychology and differential psychology. A major objective of cognitive psychology has been an accounting of learning or performance in specific tasks, i.e., within domain phenomena. The results for the most part have been a variety of microtheories, each optimized for the phenomenon at hand, just as different microtheories of item difficulty are likely to emerge from attempts to implement RGM. Even if the microtheories are successful there is another aspect of the data that must be accounted for, namely interdomain covariation.

An accounting of inter-domain covariation is really not different from the "transfer problem" that has persisted in learning and cognitive psychology. Indeed, Messick (1972) has proposed the transfer problem as an arena for incorporating function into individual differences theorizing. Whereas psychometricians have attempted to account for the degree to which test scores covary--and for the most have failed, according to Carroll (1988)--for the cognitive psychologist the problem is to account for transfer--or more often than not lack thereof. Psychometricians may have described the covariation among a wide range of tests but such descriptions do not constitute an accounting. Similarly, cognitive

psychologists are often at a loss to explain lack of transfer According to Larkin (1989, p. 303): "Although attractive, the notion that transferable knowledge is a core of general problem-solving skills has been historically unproductive." She argues that the answer lies in incorporating more detail:

> Instruction in skills is most effective if we can understand in detail what we want to teach and focus instruction accordingly. Detailed models of strategies for related domains, methods for setting subgoals, knowledge of task management, and learning skills seem a promising road to this end (Larkin, 1989, p. 304).

Knowing that cognitive and differential psychology share concerns is reassuring but does not answer the question whether a single framework can serve as the foundation for RGM across a variety of domains. A similar question has been raised by computational psychologists (Boden, 1988, p. 171) who phrase the questions in terms of a general theory of problem solving, and by intelligence theorist (e.g., Sternberg & Powell, 1982).

One answer, of course, is that such a general theory is not possible, a view taken by modularity psychologists (e.g., Fodor, 1983) and by cognitive anthropologists who argue that the modelling of problem solving must take context and situations into account (Lave, 1988). Others, however, argue that it is indeed possible, and they propose a scheme, or architecture, under which we can subsume a variety of problem solving behaviors. (Newell, 1989) The Newell-Simon (1972) approach to problem solving is especially relevant to psychometrics because of its concern with problem difficulty. As early as 1972 Newell and Simon (1972, p. 93) discussed at length problem difficulty in ways that are totally consistent with the componential approaches to psychometric modelling of Carroll (1976), Sternberg (1977) and Whitely (1980) and even the disjunctive-conjunctive distinction discussed by Jannarone (in preparation). In the Newell-Simon theory, the problem solver is viewed as constructing problem spaces for each problem. The difficulty of a problem is then, in part, a function of the problem space:

> "The size of the problem space provides one key to the estimation of problem difficulty. The problem space defines the set of possibilities for the solution, *as seen by the problem solver.* (Newell & Simon, 1972, p. 93, italics added).

Clearly, Newell and Simon had an idiographic view of difficulty in mind when they defined

problem spaces as being specific to a problem solver, but later on in the book they consider nomothetic individual differences and attribute them primarily to the contents of long term memory and, to "basic structure" (p. 865):

> ...it follows that any proposal for communality among problem solvers not attributable to basic structure must be represented as an identity or similarity in the contents of the LTM--in the production system or in other memory structures (p. 865).

The applicability of the Newell-Simon framework to an accounting of individual differences on a psychometric instrument, the Raven Progressive Matrices, has been demonstrated by Carpenter, Just and Shell (in press). They account for performance on this test, considered to be one of the purest measures of intelligence, by explicating the differences in level of performance in the form of simulation models that perform at different levels. Briefly, the kinds of models they postulate consist of a set of productions, or condition-action rules, to represent the content of long term memory. When those productions are activated by the requirements of the problem they deposit information in short term memory. The solution to a problem is obtained by operating on the content of short term memory. Within this framework individual differences can be a function of the content of long term memory and the working memory capacity, or "basic structure" as originally formulated by Newell and Simon. But in the case of the Raven, which uses totally novel stimuli, working memory capacity may in fact be more important because there is not much information to be retrieved from long term memory.

In general, and especially with achievement tests, long term memory would be expected to play a larger role. However, "basic structure," or working memory capacity, would seem to be centrally involved, even in domains that are knowledge dependent, because working memory capacity is involved not only in the solution of the current problem but was also involved in the creation and storing of the knowledge which is now triggered to solve the current problem. Thus, working memory capacity may be the equivalent of *g* in differential psychology, postulated by Spearman (1923) to account for the consistent covariation among intellectual tasks. However, we now know that there is more than *g*. A break down of the "factorial pie" in terms of crystallized and fluid intelligences (e.g., Horn, 1970) has received a wide acceptance (e.g., Snow & Lohman, 1989). This breakdown seems to fit with an

equating fluid intelligence to working memory capacity, and crystallized intelligence to productions, or knowledge.

The notion that there can be an all-encompassing theory of problem solving has not gone unchallenged (e.g., Boden, 1988, p. 171). One argument is that problem solving is not computationally, encapsulated, but involves *cognitive penetrable* phenomena, a term originated by Pylyshyn (1984), which means that the problem solver is influenced by his or her desires and beliefs. This view would seem to suggest that the actual difficulty of a problem for a given individual would be a function of that person's ability, the nature of the problem, and its desires and his or her beliefs. From a psychometric perspective this need not be a fatal a problem as it might be to a purely psychological theory because psychometric models can deal with error. Moreover, there is no reason why the penetrability could not itself be modeled by establishing the link of beliefs and desires into a response mechanism (cf Boden, p. 174). An example of modelling penetrability within a psychometric framework is provided by Colberg and Nester (1987), who are able to anticipate the range of illogical beliefs and incorporate those as part of the prediction of difficulty of deductive reasoning items. In short, penetrability need not be a fatal problem, at least from a psychometric perspective.

The Newell-Simon approach has been characterized as embodying a symbolic paradigm (Smolensky, 1986). A contender to the Newell-Simon framework argues for a subsymbolic approach. Smolensky 1986, for example illustrates electronic problem solving from a subsysmbolic perspective where instead of of representing knowledge as productions, knowledge is distributed in a network the nodes of which represent bits of knowledge. The states of that network are assumed to correspond to psychological meaningful states.

Both symbolic and subsymbolic approaches to modelling cognition lend themselves to psychometric modelling, and are appealing because of their psychological underpinnings but seem better suited for within-task analyses. The covariation among tasks needs also to be accounted for. Such an accounting could come about from a detailed analysis of studies that describe performance covariation across a variety of tasks. The most obvious source of data for such an analysis is found in the factor analytic literature. The value of such analyses is demonstrated by two metanalyses. Snow,

Kyllonen and Marshalek (1984) reanalyzed several data sets and concluded that Guttman's (1954) radex theory of intelligence was correct. That is, performance across a variety of cognitive tasks can be described as a circular map. Located at the center of the map we find performance on the Raven's Progressive Matrix test, presumably representing *g*. Moreover, the circle can be divided into three slices corresponding to verbal, quantitative and spatial domains. The tests on the periphery are simpler, and as we move toward the center their complexity increases. The rich detail provided by Snow at al. seems to be beyond the scope of the Newell-Simon or mental models frameworks. The second reanalysis of existing data was provided by Carroll (e.g., 1980) who postulates ten basic information processing components as the basis for the factors that factor analysts have postulated to account for covariation among test scores.

Clearly, we are not at the point where we decide what is the best approach to a general psychological framework for test construction. Perhaps, a variety of perspectives should be encouraged. What RGM does is to provide a Popperian mechanism for psychometric modeling. According to Popper (1959) the scientific status of a theory depends on its falsifiability. Moreover, evidence in favor of a theory is not as convincing unless that evidence was obtained as part of a challenge, i.e., in an attempt to falsify the theory. RGM links item construction and response modeling in a single package so that the linkage, i.e., the predictions about response behavior, are challenged every time a test is administered. Thus, the administration of a test becomes a psychological experiment, which in turn may lead to the improvement of both theories and tests.

Acknowledgements

I'm grateful to Irving Sigel, Larry Frase, Norman Frederiksen Robert Mislevy and Lawrence Stricker for valuable comments on earlier versions of this manuscript. Those suggestions I did not incorporate into this manuscript will surely improve future ones.

References

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. Applied Psychological Measurement, 12, 117-128.

Anderson, J. R. The adaptive character of thought. Hillsdale, NJ: Erlbaum

Baker, F. B. (1988). Computer technology in test construction and processing. In R. L. Linn (Ed.), Educational Measurement (pp. 409-428). New York: Macmillan.

Bejar, I. I. (in press). A generative analysis of a three-dimensional spatial task. Applied Psychological Measurement.

Bejar, I. I. (in preparation). Leveraging the computer for test delivery by automating the scoring of open-ended items. Princeton, NJ: Educational Testing Service.

Bejar, I. I. (1983). Achievement testing: Recent advances. Beverly Hills, CA: Sage Publications.

Bejar, I. I. (1984). Educational diagnostic assessment. Journal of Educational Measurement, 21, 175-189.

Bejar, I. I. (1988). A sentence-based automated approach to the assessment of writing: A feasibility study. Machine-Mediated Learning, 2, 321-332.

Bejar, I. I. (1989, August). Generative response modeling and test development job aids: An approach to improving validity through technology. Presented in Implications of Measurement Theory for Language Assessment at the ETS TOEFL Invitational Symposium on Language Acquisition and Language Assessment. Princeton, NJ.

Bejar, I. I., Chaffin, R., & Embretson, S. (in press). Cognitive and psychometric analysis of analogical problem solving. New York: Springer-Verlag.

Bejar, I. I., Stabler, E. P., & Camp, R. (1987). Syntactic complexity and psychometric difficulty: A preliminary investigation (RR-87-25). Princeton, NJ: Educational Testing Service.

Bejar, I. I., & Yocom, P. (in press). A generative approach to the modeling of hidden-figure items. Applied Psychological Measurement.

Bock, R. D., Gibbon, R., & Muraki, E. (1988). Full-information item factor analysis. Applied Psychological Measurement, 12, 261-280.

Boden, M. A. (1988). Computer models of mind: Computational approaches in theoretical psychology. New York: Cambridge University Press.

Borko, H., & Livingston, C. (1989). Cognition and improvisation: Differences in mathematics instruction by expert and novice teachers. American Educational Research Journal, 26, 473-498.

Bormuth, J. R. (1970). On the theory of achievement test items. Chicago: University of Chicago Press.

Braun, H., Carlson, S., & Bejar, I. I. (1989). Psychometric foundations of testing based on patient management problems (RM-89-2). Princeton, NJ: Educational Testing Service.

Breland, H. M., & Lytle, E. G. (1990) Computer-assisted writing assessment using WordMAP(TM). Presented at NCME 1990 Annual Conference in Paper Presentation "Computer-Assisted Assessment: Innovations and Investigations," chaired by M. Kinzie. Boston, MA.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 2, 155-192.

Burton, R. R. (1982). Diagnosing bugs in a simple procedural skill. In D. Sleeman, & J. S. Brown (Eds.), Intelligent tutoring systems (pp. 157-183). New York: Academic Press.

Butterfield, E. C., Nielsen, D., Tangen, K. L., & Richardson, M. B. (1985). Theoretically based psychometric measures of inductive reasoning. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 77-147). New York: Academic Press.

Byrd, R. J., Calzolari, N., Chodorow, M. S., Edwards, D., Klavans, J. L., & Neff, M. S. (1987). Tools and methods for compututional linguistics. Computational Linguistics, 13, 219-240.

Campbell, A. C. (1961). Some determinants of the difficulty of non-verbal classification items. Educational and Psychological Measurement, 21, 899-913.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C. P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. American

Educational Research Journal, 26, 499-531.

Carpenter, P. A., Just, M. A., & Shell, P. (in press). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. Psychological Review.

Carroll, J. B. (1972). Stalking the wayward factors. Contemporary Psychology, 17, 321-324.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new "Structure of Intellect." In L. Resnick (Ed.), The nature of intelligence (pp. 27-56). Hillsdale, NJ: Erlbaum

Carroll, J. B. (1980). Measurement of abilities constructs. In Construct Validity in Psychological Measurement. Proceedings of a colloquium on theory and application in education and employment. Princeton, NJ: Educational Testing Service.

Carroll, J. B. (1980). Individual difference relations in psychometric and experimental cognitive tasks. L. L. Thurstone Psychometric Laboratory Report No. 163 (ERIC Doc ED-191-891)

Carroll, J. B. (1988, April). Factor analysis since Spearman: Where do we stand? What do we know? Presented at a symposium, Learning and Individual Differences: Abilities, Motivation, and Methodology, Department of Psychology, University of Minnesota.

Chaffin, R., & Herrmann, D. J. (1987).Relation element theory: A new account of the representation and processing of semantic relations. In D. Gorfein & R. Hoffman (Eds.), Memory and learning: The Ebbinghaus centennial conference. Hillsdale, NJ: Erlbaum.

Chi, M. T. H., Feltovich, P. J., Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MIT Press.

Colberg, M., & Nester, M. (1987, August). The use of illogical biases in psychometrics. Paper presented at the 7th International Congress of Logic, Methodology and Philosophy of Science. Moscow, USSR.

Cole, N. S. (1990). Conceptions of educational achievement. Educational Researcher, 19, 2-7.

Corballis, M. C. (1982). Mental rotation: Analysis of a paradigm. In M. Potegal (Ed.), Spatial

abilities: Developmental and psychological foundations. New York: Academic Press.

Cronbach, L. J. (1970). Review of On the theory of achievement test items, by J. R. Bormuth. Psychometrika, 35, 509-511.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Curtis, M. E. (1987). Vocabulary testing and vocabulary instruction. In M. G. McKeown, & M. E. Curtis (Eds.), The nature of vocabulary acquisition (pp. 37-51). Hillsdale, NJ: Erlbaum.

Davies, A. D. M., & Davies, M. G. (1965). The difficulty and graded scoring of Elithorn's perceptual maze test. British Journal of Psychology, 14, 295-302.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. Psychological Review, 96, 433-458.

Egan, D. E. (1979). Testing based on understanding: Implications from studies of spatial ability. Intelligence, 3, 1-15.

Ekstrom, R. B., French, J., & Harman, H. (1976). Kit of Factor-Referenced Cognitive Tests. Princeton, NJ: Educational Testing Service.

Elithorn, A., Jones, D., Kerr, M., & Lee, D. (1964). The effects of the variation of two physical parameters on empirical difficulty in a perceptual maze test. British Journal of Psychology, 55, 31-37.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 175-197.

Fellbaum, C. (1987). A preliminary analysis of cognitive-linguistic aspects of sentence completion tasks. In R. O. Freedle, & R. P. Duran (Eds.), Cognitive and linguistic analyses of test performance (pp. 193-207), Vol. XXII in the series Advances in Discourse Processes, R. O. Freedle (Ed.). Norwood, NJ: Ablex.

Fodor, J. (1983). The modularity of mind. Cambridge, MA: MIT Press.

Fischer, G. H. (1973). The linear logist test model as an instrument in educational research. Acta Psychologica, 37, 359-374.

Frederiksen, N. (1990). Introduction. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition. Hillsdale, NJ: Erlbaum.

Frederiksen, N. (1986). Construct validity and construct similarity: Methods for use in test development and test validation. Multivariate Behavioral Research, 21, 3-28.

Fulton, S. L., & Pepe, C. O. (1990 ,January). An introduction to model-based reasoning. AI Expert, 48-55.

Gardner, H. (1985). The mind's new science: A history of the cognitive revalution. New York: Basic Books.

Galotti, K. M. (1989). Approaches to studying formal and everyday reasoning. Psychological Bulletin, 105, 331-351.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.

Glaser, R. (1988). Cognitive and environmental perspectives on assessing achievement. In Assessment in the Service of Learning: Proceedings of the 1987 ETS Invitational Conference (pp. 37-43). Princeton, NJ: Educational Testing Service.

Goldin, G. A., & McClintock, C. E. (Eds.). (1984). Task variables in mathematical problem solving. Philadelphia, PA: Franklin Institute Press.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. British Journal of Mathematical and Statistical Psychology, 42, 139-167.

Green, D. R. (Ed.). (1974). The aptitude-achievement distinction: Proceedings of the Second CTB/McGraw-Hill Conference on Issues in Educational Measurement. Monterey, CA: CTB/McGraw-Hill.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Guttman, L. A. Measurement as structural theory. Psychometrika, 26, 329-347

Guttman, L. A. (1954). A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), Mathematical thinking in the social sciences. Glencoe, IL.

Guttman, L. A. (1969). Integration of test design and amalysis. In Proceedings of the 1969 Invitational

Conference on Testing Problems. Princeton NJ: Educational Testing Service.

Guttman, L. A. (1980) Integration of test desig and analysis: Status in 1979. In W. B. Schrader (Ed.) Measuring Achievement: Progress Over a Decade. Proceedings of the 1979 ETS Invitational Conference. San Francisco: Jossey-Bass Inc.

Hall, R., Kibler, D., Wenger, E., & Truxaw, C. (1989). Exploring the episodic structure of algebra story problem solving. Cognition and Instruction, 6, 223-283.

Harmon, P. (1986). Intelligent job aids: How AI will change training in the next five years. In G. Kearsley (Ed.), Artificial intelligence and instruction: Application and methods. Reading, MA: Addison-Wesley.

Hively, W. (1974). Introduction to domain referenced testing. Educational Technology, 14, 5-9.

Hively, W., Paterson, H. L., & Page, S. H. (1968). A universe-defined system of arithmetic tests. Journal of Educational Measurement, 5, 275-290.

Horn, J. L. (1970). Organization of data on life span development of human abilities. In L. R. Goulet & P. B. Baltes (Eds.), Life-span developmental psychology: research and theory (pp. 423-466). New York: Academic Press.

Irvine, S. H., Dunn, P. L., & Anderson, J. D. (1989). Towards a theory of algorithm-determined cognitive test construction (Report). Devon, UK: Polytechnic South West.

Jannarone, R. (in preparation). Measuring quickness and correctness concurrently: A cognitive IRT approach. Columbia, SC: Psychology Department, University of South Carolina.

Jarjoura, D., & Brennan, R. L. (1982). A variance components model for measurement procedures associated with a table of specifications. Applied Psychological Measurement, 6, 161-171.

Jensen, A. R. (1980). Bias in mental testing. New York: Free Press.

Johnson-Laird, P. N. (1983). Mental models: Towards a cognitive science of language, inference, and consciousness. Cambridge, MA: Harvard University Press.

Johnson, Laird, P. N. (1988). A taxonomy of thinking. In R. J. Sternberg & E. E. Smith (Eds.), The psychology of human thought. New York: Cambridge University Press.

Johnson-Laird, P. N., Byrne, R. M. J., & Tabossi, P. (1989). Reasoning by model: The case of

multiple quantification. Psychological Review, 96, 658-673.

Just, M. A., & Carpenter, P. A. (1987). The psychology of reading and language comprehension. Needham, MA: Allyn and Bacon.

Katz, B. (1988). Using English for indexing and retrieving (A.I. Memo No. 1096). Cambridge, MA: Massachusetts Institute of Technology

Kieras, D. E. (1990). The role of cognitive simulation models in the development of advanced training and testing systems. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 51-73). Hillsdale, NJ: Erlbaum.

Kolen, M. J., & Harris, D. J. (1987, April). A multivariate test theory model based on item response theory and generalizability theory. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, D.C.

Kotovsky, K.,, & Simon, H. A. (1988). What makes some problems really hard: Explorations in the problem space of difficulty (ONR Report N00014-85-K-0696). Pittsburgh, PA: Community College of Allegheny County.

Kyllonen, P. C. (1990, April). Taxonomies of cognitive abilities. Presented at the American Educational Research Association Meeting. Boston, MA:.

Larkin, J. H. (1989). What kind of knowledge transfers? In L. R. Resnick (Ed.), Knowing, learning, and instruction: Essays in honor of Robert Glaser (pp. 283-305). Hillsdale, NJ: Erlbaum.

Lave, J. (1988). Cognition in practice: Mind, mathematics and culture in everyday life. New York: Cambridge University Press.

Lesgold, A., Ivill-Friel, J., & Bonar, J. (1989). Toward intelligent systems for testing. In L. R. Resnick (Ed.), Knowing, learning, and instruction: Essays in honor of Robert Glaser (pp. 337-360). Hillsdale, NJ: Erlbaum.

Loevinger, J. (1965). Person and population as psychometric concepts. Psychological Review, 72, 143-155.

Marshall, S. R. (1990, April). What students learn (and remember) from word problem instruction.

Presented at AERA 1990 Annual Convention in symposium, Penetrating to the Mathematical Structure of Word Problems, chaired by S. F. Chipman. Boston, MA.

Mayer, R. E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories and templates. Instructional Science, 10, 135-175.

Mayhew, J., & Frisby, J. (1984). Computer vision. In T. O'Shea and M. Eisenstadt (Ed.), Artificial intelligence: Tools, techniques, and applications. New York: Harper & Row.

Merwin, J. C. (1977). Considerations in exploring alternatives to standardized tests. In A. J. Nitko (Ed.), Exploring alternatives to current standardized tests: Proceedings of the 1976 National Testing Conference (pp. 5-24). Pittsburgh, PA: University of Pittsburgh.

Messick, S. (1972). Beyond structure: In search of functional models of psychological process. Psychometrika, 37, 357-375.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. Psychological Bulletin, 89, 575-588.

Miller, G. A. (1962). Some psychological studies of grammars. American Psychologist, 17, 748-762.

Miller, G. A., Fellbaum, C., Kegl, J., & Miller, K. (1988) WORDNET: An electronic lexical reference system based on theories of lexical memory (Report No. 29). Princeton, NJ: Cognitive Science Laboratory, Princeton University.

Miller, P. L. (1984). A critiquing approach to expert computer advice: Attending. Palo Alto, CA: Kaufmann.

Mislevy, R. [this volume]

Mislevy, R. J., & Sheehan, K. M. (1990, June). How to equate tests with little or no data. Presented at the Psychometric Society Meeting. Princeton, NJ:.

Newell, A. (1989). Putting it all together. In D. R. Lahr & K. Kotovsky (Eds.), Complex information processing: The inpact of Herbert A. Simon (pp. 339-445) Hillsdale, NJ: Erlbaum

Newell, A., & H. A. Simon (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.

Osburn, H. G. (1968). Item sampling for achievement testing. Educational and Psychological Measurement, 28, 95-104.

Parker, R. C., & Miller, R. A. (1988). Using causal knowledge to create simulated patient cases: CPCS Project as an extension of INTERNIST-1. In P. L. Miller (Ed.), Selected topics in medical artificial intelligence (pp. 99-115). New York: Springer-Verlag.

Pearl, J. (1987). Probabilistic reasoning in intelligent systems: Networks of plausible inference. San Mateo, CA: Kaufmann.

Pellegrino, J. W., & Glaser, R. (1982). Analyzing aptitudes for learning: Inductive reasoning. In R. Glaser (Ed.), Advances in instructional psychology (Vol. 2). Hillsdale, NJ: Erlbaum.

Perfetti, C. A. (1989). There are generalized abilities and one of them is reading. In L. R. Resnick (Ed.), Knowing, learning, and instructions: Essays in honor of Robert Glaser (pp. 307-335). Hillsdale, NJ: Erlbaum.

Popper, Sir Karl. (1959) The logic of scientific discovery. Hutchinson & Co.

Prince, A., & Pinker, S. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Cognition, 28.

Pylyshyn, Z. W. (1984). Computation and cognition: Toward a foundation for cognitive science. Cambridge, MA: MIT Press.

Ranney, M., & Thagard, P. (1988). Explanatory coherence and belief revision in naive physics (Technical Report No. UPITT/LRDC/ONR/APS-17). Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Reed, S. K., Ackinclose, C. C., & Voss, A. A. (1990). Selecting analogous problems: Similarity versus inclusiveness. Memory and Cognition, 18, 83-98.

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983) Development of children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.), The development of mathematical thinking (pp. 153-196). New York: Academic Press.

Roid, G., & Haladyna, T. (1982). A technology for test-item writing. New York: Academic.

Rummelhart, D. E., McClelland, J. L., & The PDR Research Group. (1986). Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation. Cambridge, MA: MIT Press.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. Science, 171, 701-703.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. Harvard Educational Review, 57, 1-22.

Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. Psychological Review, 70, 534-546.

Skinner, B. F. (1957). Verbal behavior. New York: Appleton-Century Crafts.

Smolensky, P. Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland and the PDP Research Group (Eds.) Parallel distributed processing: Volume 1: Foundations (pp. 194-281). Cambridge MA: MIT Press

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 2). Hillsdale, NJ: Erlbaum.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 263-331). New York: Macmillan.

Spearman, C. (1923). The nature of 'intelligence' and the principles of cognition. New York: Macmillan.

Sternberg, R. J. (1977). Intelligence, information processing and analogical reasoning: The componential analysis of human abilities. New York: Wiley.

Sternberg, R. J. (1987). Most vocabulary is learned from context. In M. G. McKeown, & M. E. Curtis (Eds.), The nature of vocabulary acquisition (pp. 89-105). Hillsdale, NJ: Erlbaum.

Sternberg, R. J., & Powell, J. S. (1982). Theories of intelligence. In R. J. Sternberg (Ed.), Handbook of human intelligence (pp. 975-1005). New York: Cambridge University Press.

Tate, M. W. (1948). Individual differences in speed of response in mental test materials of varying degrees of difficulty. Educational and Psychological Measurement, 8, 353-374.

Tatsuoka, K. K., & Tatsuoka, M. M. (1987). Bug distribution and statistical pattern classification.

Psychometrika, 52, 193-206.

Thagard, P. (1989). Explanatory coherence. Behavioral and Brain Sciences, 12, 435-502.

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. Psychological Bulletin, 54, 229-249.

Underwood, B. J. (1975). Individual differences as a cruicible in theory construction. American Psychologist, 30, 128-134

Uttal, W. R., Rogers, M., Hieronymous, R. & Pasich, T. (1970). Generative computer-assisted instruction in analytic geometry. Newburyport, MA: Entelek, Inc.

Warner, H. R., Haug, P., Bouhaddou, O., Lincoln, M., Warner, H. Jr., Sorenson, D., Williamson, J. W., & Fan, C. (1988, November). ILIAD as an expert consultant to teach differential diagnosis. In R. A. Greenes (Ed.), Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care (pp. 371-376). New York: Computer Society Press.

Whitely, S. E. (1980). Latent trait models in the study of intelligence. Intelligence, 4, 97-132.

Yamamoto, K. (1989). Hybrid model of IRT and latent class models (RR-89-41). Princeton, NJ: Educational Testing Service.

Zimmerman, W. S. (1954). The influence of item complexity upon the factor composition of a spatial visualization test. Educational and Psychological Measurement, 14, 106-119.

Table 1

The first three columns refer to the possible input arrangements, the last five colums refer to correct output arrangements

			Measure				
Comm-1	Comm-2	Comm-3	1	2	3	4	5
0	0	0	0	0	0	0	1
1	0	0	1	1	0	1	0
0	1	0	1	1	0	0	1
0	0	1	0	1	0	0	1
1	1	0	0	0	1	0	0
0	1	1	1	0	1	0	1
1	0	1	1	0	1	1	0
1	1	1	0	1	1	0	0

Figure Caption

Figure 1. Sample mental rotation item

Figure 2. Relationship of estimated difficulty on angular disparity at several elapsed times.

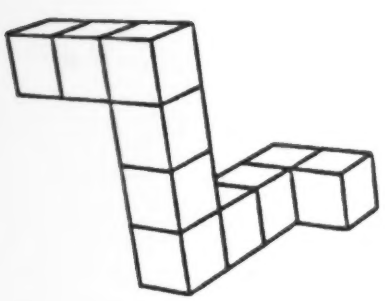
Figure 3. Typical hidden figure item and two corresponding clones.

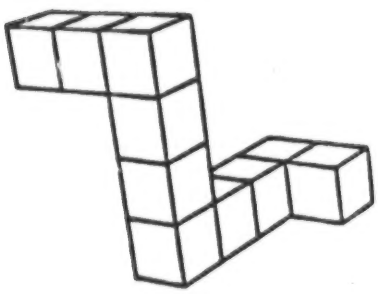
Figure 4. Regression of logit of proportion correct for pairs of clones administered to two respective random samples.

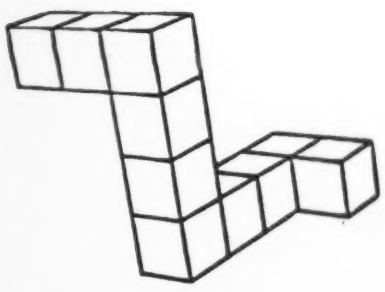
Figure 5. Cummulative response time for (a) a generating item administered to two random samples, and (b) two clones administered to respective random samples.

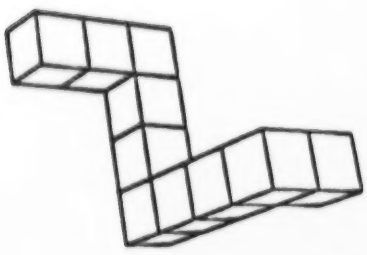
Figure 6. System for generative assessment with sentence-based items.

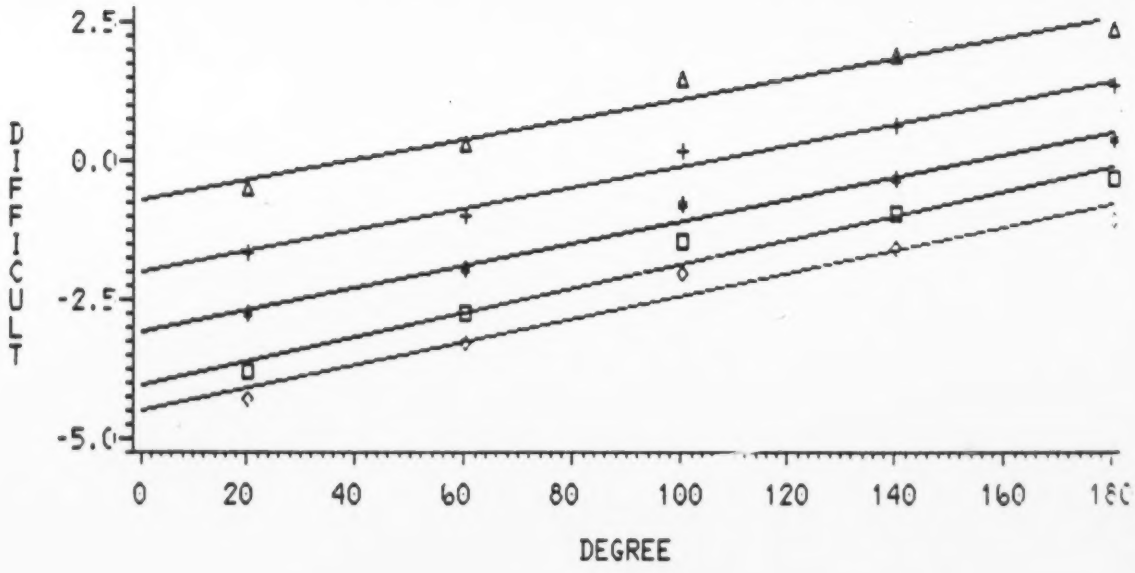
Figure 7. Electronic device.



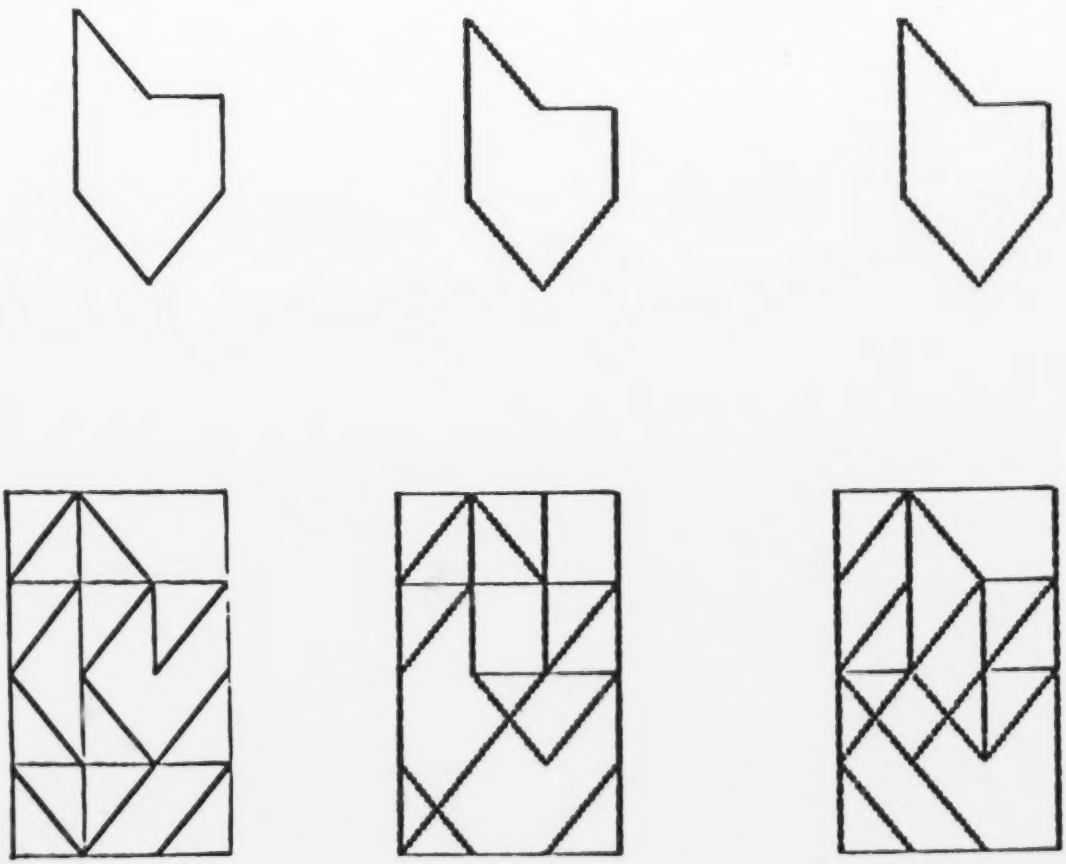


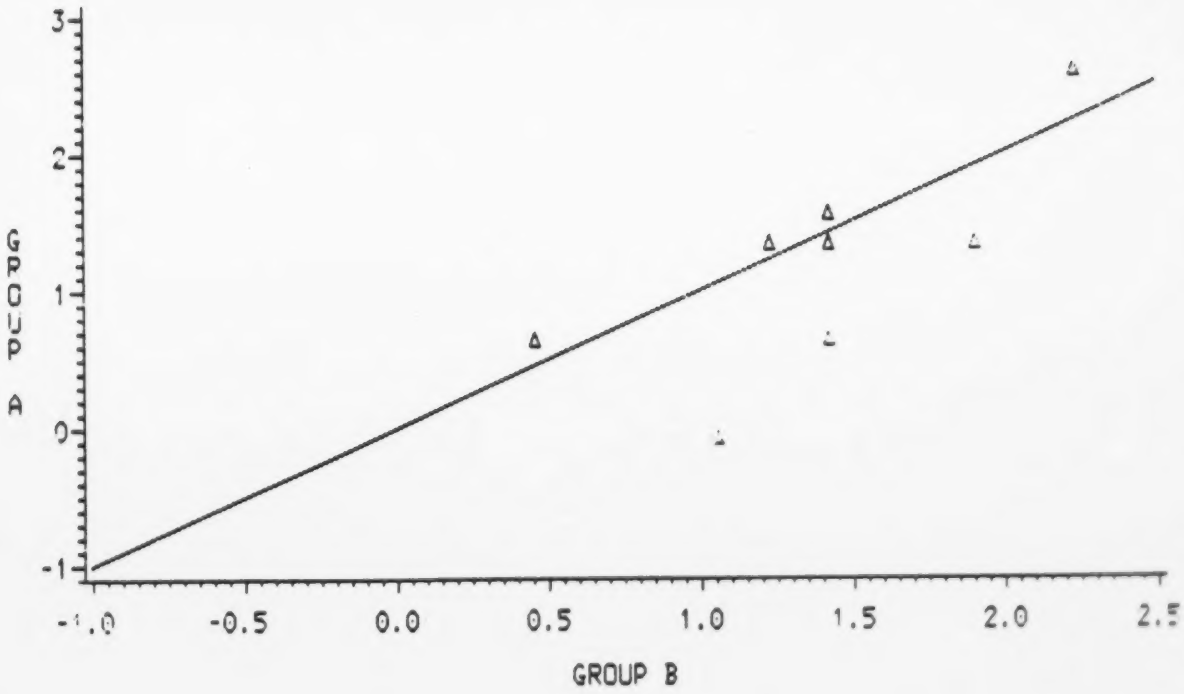




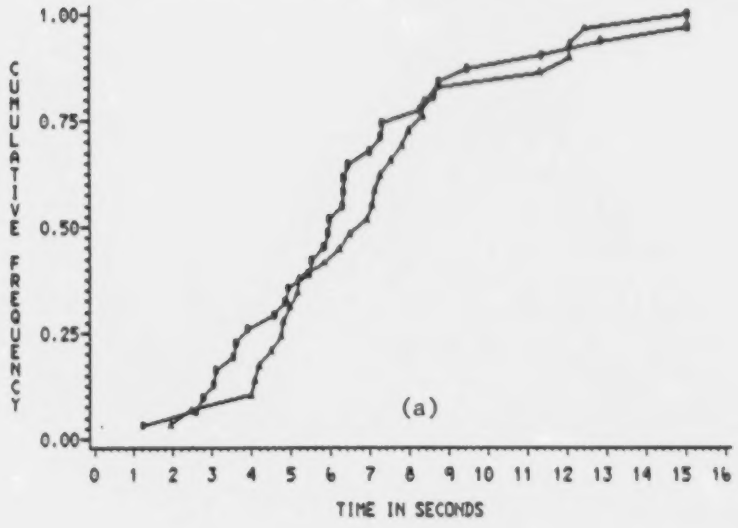


TRIANGLE= 3 SECONDS
PLUS = 4 SECONDS
STAR = 5 SECONDS
SQUARE = 6 SECONDS
DIAMOND= 7 SECONDS

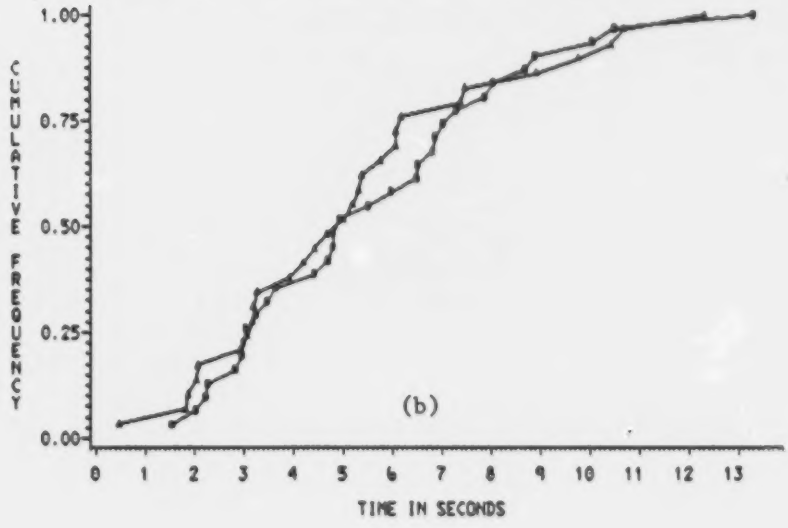

HF5
Generating Item
Close a
Close b

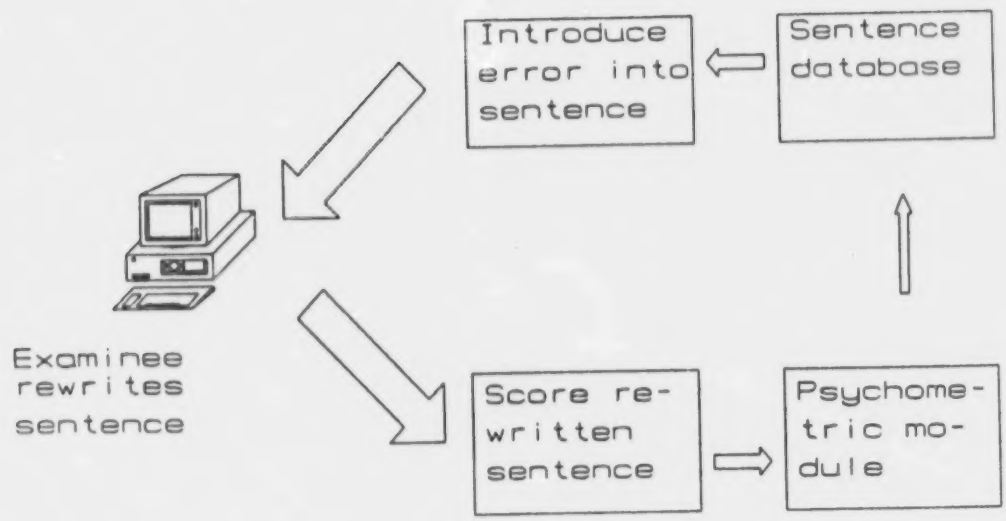

3
2
1
0
-1
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
GROUP B

DISTRIBUTION OF RESPONSE LATENCY FOR ITEM 4M

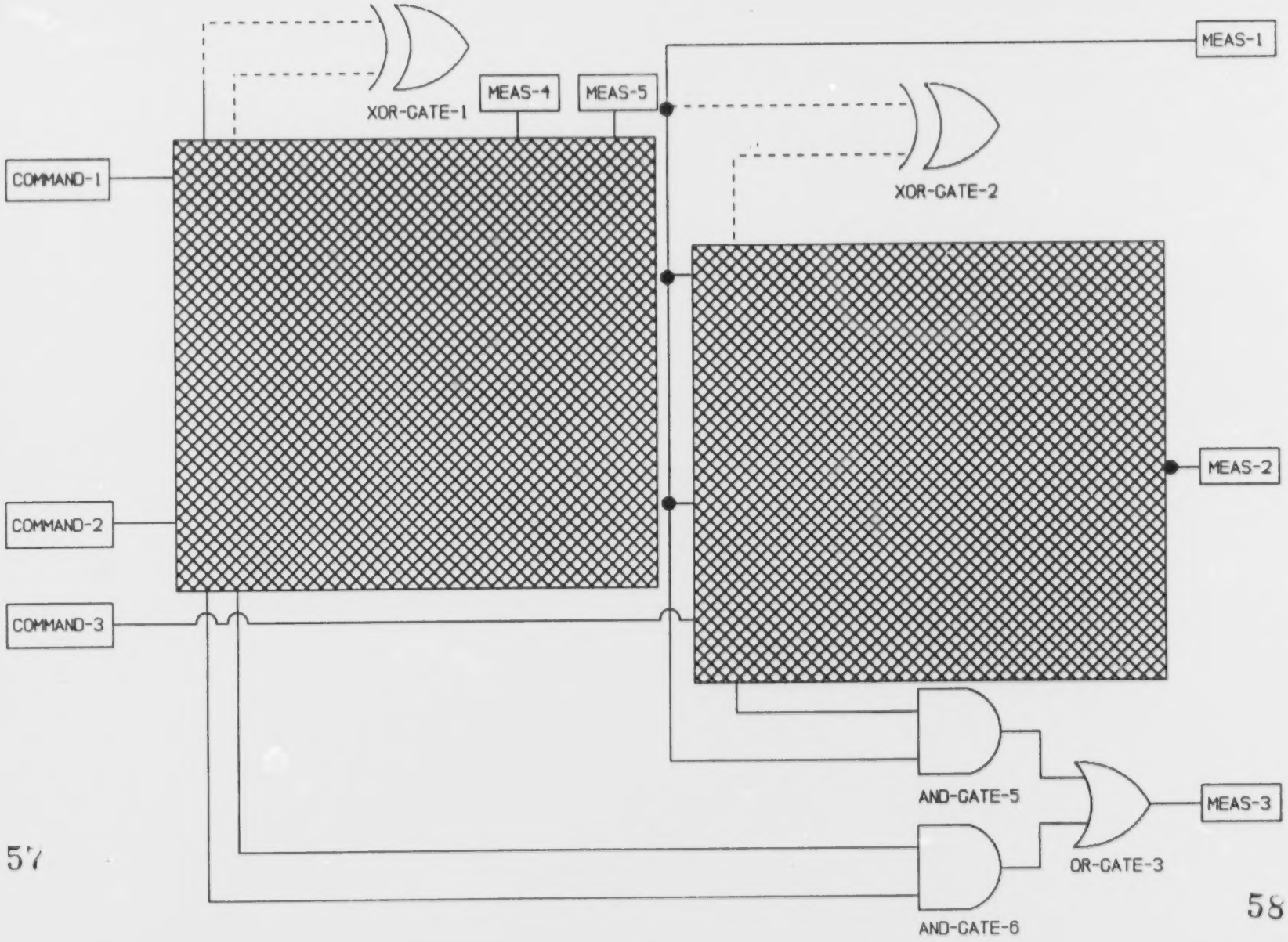


DISTRIBUTION OF RESPONSE LATENCY FOR ITEM 4C





	Original sentence without error	Original sentence with error
Examinee rewrites? yes	Is rewritten sentence semantically equivalent?	Maximum credit given if error correctly removed and none introduced
Examinee rewrites? no	Examinee given maximum credit	Note type of error not recognized by examinee


MEAS-1
XOR-GATE-1
MEAS-4
MEAS-5
XOR-GATE-2
COMMAND-1
MEAS-2
COMMAND-2
COMMAND-3
AND-GATE-5
MEAS-3
OR-GATE-3
AND-GATE-6